Uniform Crime Reporting (UCR) Program Data: A Practitioners Guide

Jacob Kaplan

2021-03-17

Contents

Preface

If you've read an article about crime or arrests in the United States in the last half century, in most cases it was referring to the FBI's Uniform Crime Reporting Program Data, otherwise known as UCR data. UCR data is, with the exception of the more detailed data that only covers murders, a *monthly number of crimes or arrests reported to a single police agency* which is then gathered by the FBI into one file that includes all reporting agencies. Think of your home town. This data will tell you how many crimes were reported for a small number of crimes or how many people (broken down by age, sex, and race) were arrested for a (larger) set of crimes in that city (if the city has multiple police agencies, it will use the agency which is the primary agency on the case, usually the local police department) in a given month. This is a very broad measure of crime, and its uses in research - or understanding crime at all - is fairly limited. Yet is has become - and will likely remain among researchers for at least the next decade - the most important crime data in the United States.

UCR data is important for three reasons:

1. The definitions are standard and all agencies follow them so you can compare across agencies and over time
2. The data is available since 1960 so there is a long period of available data
3. The data is available for most of the 18,000 police agencies in the United States so you can compare across agencies

For most of this book we'll be discussing the caveats of the above reasons - or, more directly, why these assumptions are wrong - but these are the reasons why the data is so influential.

Motivation

By the end of each chapter you should have a firm grasp on the dataset the covered and how to use it properly. However, this book can't possibly cover every potential use case for the data so make sure to carefully examine the data for your own particular use. This benefits you because you'll know your data better and become a better research because of it. This benefits me because it'll increase the quality of research in my field.

I get a lot of emails from people asking questions about this data so part of my motivation for writing this book is to create a single place that answers as many questions as I can about the data. Again, this is among the most commonly used crime datasets and there are still many current papers published with incorrect information about the data (including such simple aspects like what geographic unit data is in and what time unit it is in). So hopefully this book will decrease the amount of misconceptions about this data, increasing overall research quality.[1]

Structure of the book

This remainder of this book will be divided into eight chapters: an intro chapter briefly summarizing each dataset and going over overall issues with UCR data, six chapters each covering one of the six UCR datasets, and a final one covering county-level data, a highly flawed but common use of the UCR data. Each chapter will follow the same format: we'll start with a history of the data such as when it first because available and important changes in definitions or variables.

Next we'll discuss what the data looks like initially when you get it from the FBI - literally what it looks like in it's fixed-width ASCII format you get from the FBI and what it looks like (what each row and column mean) once it's turned into a useful format that can be read into modern software like R and Stata.[2] For most of the datasets this is a minor process but for data like

[1] Ideally, this will also decrease the number of emails I receive.

[2] To look at the data in its fixed-width ASCII format I'll use the program Notepad which can open up text files like ASCII files. For looking at the machine-readable format I'll use Stata since I think it's a bit better looking than viewing it in R. In both cases,

the arrest or homicide datasets, the conversion process is harder - and this can actually lead to changes in the resulting data. For example, in an old version of the arrest data that I released, I aggregated certain arrestee ages together since my laptop at the time couldn't handle converting data from ASCII to R and Stata without aggregating the ages (more age groups means more columns which means more computer memory needed). So anyone using that data would have less detailed data than the current dataset.

Understanding how the data moves from its rawest form (which in this case is after cleaning by the FBI) is important for being able to truly understand the data and its caveats. However, this is a fairly technical part of each chapter so feel free to skip it. Next in each chapter, we'll cover the variables included in the data and how to use them properly (including not using them at all) - this will be the bulk of each chapter. We'll end each chapter by briefly summarizing the data, how and when it's useful, and - most importantly - when you shouldn't use it.

Since manuals are boring, I'll try to include graphs and images to try to alleviate the boredom. That said, I don't think it's possible to make it too fun so sorry in advanced. This book is a mix of facts about the data, such as how many years are available, and my opinions about it, such as whether it is reliable. In cases of facts I'll just say a statement - e.g. "the offenses data is available since 1960". In cases of opinion I'll temper the statement by saying something like "in my opinion…"

Citing this book

If this data was useful in your research, please cite it. To cite this book, please use the below citation:

Kaplan J (2021). *Uniform Crime Reporting (UCR) Program Data: A Practitioners Guide*. https://github.com/jacobkap/ucrbook.

BibTeX format:

---

these will be images included in the chapter - you won't need to follow along or use either program.

Sources of UCR data

There are a few different sources of UCR data available today. First, and probably most commonly used, is the data put together by the National Archive of Criminal Justice Data (NACJD)). This a team of out of the University of Michigan who manages a huge number of criminal justice datasets and makes them available to the public. If you have any questions about crime data - UCR or other data - I highly recommend you reach out to them for answers. They have a collection of data and excellent documentation available for UCR data available on their site here. One limitation to their data, however, is that each year of data is available as an individual file meaning that you'll need to concatenate each year together into a single file. Some years also have different column names (generally minor changes like spelling robbery "rob" one year and "robb" the next) which requires more work to standardize before you could concatenate. They also only have data through 2016 which means that the most recent years (UCR data is available through 2019) of data are (as of this writing) unavailable.

Next, and most usable for the general public but limited to researchers, is the FBI's official website Crime Data Explorer. On this site you can chose an agency and see annual crime data (remember, UCR data is monthly so this isn't as detailed as it can be) for certain crimes. This is okay for the general public but only provides a fraction of the data available in the actual data so is really not good for researchers.

Finally, I have my own collection of UCR data available publicly on openICPSR, a site which allows people to submit their data for public access. For each of these datasets I've taken the raw data from the FBI (for early years of homicide data this is actually from NACJD since the FBI's raw data is wrong and can't be parsed. For later years of homicide data this is from the FBI's raw data.) and read it into R. Since the data is only available from the FBI as fixed-width ASCII files, I created a setup file (we'll explain exactly how reading in this kind of data works in the next chapter) and read the data and then very lightly cleaned the data (i.e. only removing extreme outliers like an agency having millions of arsons in a month). For each of these datasets I detail what I've done to the data and briefly summarize the data (i.e. a very short version of this book) on the data's page on openICPSR. The main advantage is that all my data has standard variable names and column names and, for data that is small

enough, provide the data as a single file that has all years. For large datasets like the arrest data I break it down into parts of the data and not all years in a single file. The downside is that I don't provide documentation other than what's on the openICPSR page and only provide data in R and Stata format. I also have a similar site to the FBI's Crime Data Explorer but with more variables available, that site is available here.

It's worth mentioning a final source of UCR information. This is the annual Crimes in the United States report released by the FBI each year around the start of October.[3] As an example, here is the website for the 2019 report. In this report is summarized data which in most cases estimates missing data and provides information about national and subnational (though rarely city-level) crime data. As with the FBI's site it is only a fraction of the true data available so is not a very useful source of crime data. Still, this is a very common source of information used by researchers.

Where to find the data used in this book

The data I am using in this book is the cleaned (we'll discuss in more detail exactly what I did to clean each dataset in the dataset's chapter but the short answer is that I did very little.) and concatenated data that I put together from the raw data that the FBI releases. That data is available on my website here. I am hosting this book through GitHub which has a maximum file size allowed that is far smaller than these data so you'll need to go to my site to download the data, it's not available through this book's GitHub repo. For some examples I'm using the data before I cleaned it of outliers (as an example of the outliers present before I removed them) so that data is not publicly available.

[3] They also release a report about the first 6-months of the most recent year of data before the October release but this is generally an estimate from a sample of agencies so is far less useful.

About the author

Jacob Kaplan holds a PhD and a master's degree in criminology from the University of Pennsylvania and a bachelor's degree in criminal justice from California State University, Sacramento. His research focuses on Crime Prevention Through Environmental Design (CPTED), specifically on the effect of outdoor lighting on crime. He is the author of several R packages that make it easier to work with data, including fastDummies and asciiSetupReader. His website allows easy analysis of crime-related data and he has released over a dozen crime data sets (primarily FBI UCR data) on openICPSR that he has compiled, cleaned, and made available to the public.

For a list of papers he has written (including working papers), please see here.

For a list of data sets he has cleaned, aggregated, and made public, please see here.

For a list of R packages he has created, please see here.

Chapter 1

An Overview of the Data

One of the first, and most important questions, I think people have about crime is a simple one: is crime going up? Answering it seems simple - you just count up all the crimes that happen in an area and see if that number of bigger than it was in previous times.

1.1 What is a crime?

1.1.1 Index crimes

Chapter 2

Offenses Known and Clearances by Arrest

The Uniform Crime Report (UCR) Program Data are an collection of agency-level crime data published by the FBI. There are a number of different data sets included in the UCR including data on crime, arrests, hate crimes, arson, and stolen property. We'll be using the Offenses Known and Clearances by Arrest data set (the "crime" data set), which is the most commonly used data set in the UCR and is sometimes used as a shorthand for UCR data. In this lesson we'll use "UCR" and "Offenses Known and Clearances by Arrest" interchangeably but keep in mind that doing so is technically incorrect.

You can read more about the UCR program and all of the data sets it includes on the National Archive of Criminal Justice Data (NACJD) page here. You can also check out my site Crime Data Tool which visualizes several of the UCR data sets and has info in the FAQ page explaining the data.

Nearly every police agency in the United States - approximately 18,000 agencies - now report their data to the FBI which compiles and releases the UCR data (some states have their agencies report to a state department which then sends the data to the FBI). This data is available since 1960 though early years have many fewer agencies reporting than do so in later years.

The data file has annual data on the number of crimes reported, the number of crimes cleared, the number cleared where all offenders are under age 18, and the number of unfounded crimes. We'll discuss each of these a bit further

as we dive into the data. Agencies report the monthly number of each crime though the data we'll work with has aggregated that to annual counts.

Due to its longevity (it has data since 1960) and ubiquity (almost every agency reports and has done so for many years) it is a popular data set for criminologists.

2.1 Exploring the UCR data

We are going to look at data with the combined annual count of crimes for every year available - 1960-2017 - which I've made available here. The FBI releases the data as a single file per year and each file has monthly counts of crime. This data set does some cleaning for us by aggregating yearly and making it a single file for the whole time period. The first step when working with this UCR data is loading it into R. As with loading any data, make sure that your working directory path is correctly set using `setwd()` so R knows which folder the data is in.

We can see this is a very big file - 159 columns and nearly a million rows! Normally we'd use the `head()` function to see the first 6 rows of every column but since this data has so many columns we won't do that as it'd be hard to read. Instead we can use `View()` to open what's essentially an Excel file showing our data. This is useful to quickly glance at the data but is limited as it can bias us to believe that the first several rows are representative of the data (an issue also present with `head()`). But, for a first glance it is useful and will be supplemented by better checks below.

From looking at the data in `View()` we can see that the units are agency-years. Each row is a single agency for a single year. This is useful because it tells us we will have crime in agencies over time, which is a very common unit of crime data. Let's take a look at how many agencies report each year using the `table()` function which says how many times each value occurs for the column we select. This is also a useful check on if every year from 1960 to 2017 is actually available - don't just trust that the data has what it says it has!

From these results it's clear that there are huge differences in how many agencies report in early years compared to more recent years. Is this an issue in an analysis? From the above table it is concerning but not entirely

clear there is an issue depending on our specific analysis. It we only care about recent years then it wouldn't matter. If we only use large agencies, then knowing that relatively few agencies reported in 1960 doesn't mean that few large agencies reported. For that you'd have to look closer at only the agencies you want to study - we won't do that here but keep it in mind.

2.2 ORIs - Unique agency identifiers

In the UCR and other FBI data sets, agencies are identified using **OR**iginating Agency **I**dentifiers or ORIs. These are unique ID codes used to identify an agency. If we used the agency's name we'd end up with some duplicates For example, if you looked for the Philadelphia Police Department using the agency name, you'd find both the "Philadelphia Police Department" in Pennsylvania and the one in Mississippi.

Each ORI is a 7-digit value starting with the state abbreviation (for some reason the FBI incorrectly puts the abbreviation for Nebraska as NB instead of NE) followed by 5 numbers. In the NIBRS data (another FBI data set) the ORI uses a 9-digit code - expanding the 5 numbers to 7 numbers. When dealing with specific agencies, make sure to use the ORI rather than the agency name to avoid any mistakes.

For an easy way to find the ORI number of an agency, use this site. Type an agency name or an ORI code into the search section and it will return everything that is a match.

2.3 Hierarchy Rule

The UCR has what is called the Hierarchy Rule where only the most serious crime in an incident is reported (except for motor vehicle theft which is always included). For example if there is an incident where the victim is robbed and then murdered, only the murder is counted as it is considered more serious than the robbery.

How much does this affect our data in practice? Actually very little. Though the Hierarchy Rule does mean this data is an under-count, data from other sources indicate that it isn't much of an under count. The FBI's other data

set, the National Incident-Based Reporting System (NIBRS) contains every crime that occurs in an incident (i.e. it doesn't use the Hierarchy Rule). Using this we can measure how many crimes the Hierarchy Rule excludes (Most major cities do not report to NIBRS so what we find in NIBRS may not apply to them). In over 90% of incidents, only one crime is committed. Additionally, when people talk about "crime" they usually mean murder which, while incomplete to discuss crime, means the UCR data here is accurate on that measure.

2.4 Which crimes are included?

If you look back at the output when we ran `names(offenses_known_yearly_1960_2017)` you'll see that it produced five broad categories of columns. The first was information about the agency including population and geographic info, then came four columns with the same values except starting with "actual", "tot_clr", "clr_18", and "unfound". Following these starting values were 30 crime categories. We'll discuss what each of those starting values mean in a bit, let's first talk about which crimes are included and what that means for research.

2.4.1 Index Crimes

The Offenses Known and Clearances by Arrest data set contains information on the number of "Index Crimes" (sometimes called Part I crimes) reported to each agency. These index crimes are a collection of eight crimes that, for historical reasons based largely by perceived importance in the 1920's when the UCR program was first developed, are used as the primary measure of crime today. Other data sets in the UCR, such as the Arrests by Age, Sex, and Race data and the Hate Crime data have more crimes reported.

The crimes are, in order by the Hierarchy Rule -

1. Homicide
    - Murder and non-negligent manslaughter

- Manslaughter by negligence

2. Rape
   - Rape
   - Attempted rape
3. Robbery
   - With a firearm
   - With a knife of cutting instrument
   - With a dangerous weapon not otherwise specified
   - Unarmed - using hands, fists, feet, etc.
4. Aggravated Assault (assault with a weapon or causing serious bodily injury)
   - With a firearm
   - With a knife of cutting instrument
   - With a dangerous weapon not otherwise specified
   - Unarmed - using hands, fists, feet, etc.
5. Burglary
   - With forcible entry
   - Without forcible entry
   - Attempted burglary with forcible entry

6. Theft (other than of a motor vehicle)
7. Motor Vehicle Theft
    - Cars
    - Trucks and buses
    - Other vehicles
8. Arson
9. Simple Assault

For a full definition of each of the index crimes see the FBI's Offense Definitions page here.

Arson is considered an index crime but is not reported in this data - you need to use the separate Arson data set of the UCR to get access to arson counts. The ninth crime on that list, simple assault, is not considered an index crime but is nevertheless included in this data.

Each of the crimes in the list above, and their subcategories, are included in the UCR data. In most reports, however, you'll see them reported as the total number of index crimes, summing up categories 1-7 and reporting that as "crime". These index crimes are often divided into violent index crimes - murder, rape, robbery, and aggravated assault - and property index crimes - burglary, theft, motor vehicle theft.

2.4.2 The problem with using index crimes

The biggest problem with index crimes is that it is simply the sum of 8 (or 7 since arson data usually isn't available) crimes. Index crimes have a huge range in their seriousness - it includes both murder and theft. This is clearly wrong as 100 murders is more serious than 100 thefts. This is especially a problem as less serious crimes (theft mostly) are far more common than more serious crimes (in 2017 there were 1.25 million violent index crimes in the United States. That same year had 5.5 million thefts.). So index crimes

under-count the seriousness of crimes. Looking at total index crimes is, in effect, mostly just looking at theft.

This is especially a problem because it hides trends in violent crimes. San Francisco, as an example, has had a huge increase in index crimes in the last several years. When looking closer, that increase is driven almost entirely by the near doubling of theft since 2011. During the same years, violent crime has stayed fairly steady. So the city isn't getting more dangerous but it appears like it is due to just looking at total index crimes.

Many researchers divide index crimes into violent and nonviolent categories, which helps but is still not entirely sufficient. Take Chicago as an example. It is a city infamous for its large number of murders. But as a fraction of index crimes, Chicago has a rounding error worth of murders. Their 653 murders in 2017 is only 0.5% of total index crimes. For violent index crimes, murder makes up 2.2%. What this means is that changes in murder are very difficult to detect. If Chicago had no murders this year, but a less serious crime (such as theft) increased slightly, we couldn't tell from looking at the number of index crimes.

2.4.3 Rape definition change

Starting in 2013, rape has a new, broader definition in the UCR to include oral and anal penetration (by a body part or object) and to allow men to be victims. The previous definition included only forcible intercourse against a woman. As this revised definition is broader than the original one post-2013, rape data is not comparable to pre-2013 data.

2.5 Actual offenses, clearances, and unfounded offenses

For each crime we have four different categories indicating the number of crimes actually committed, the number cleared, and the number determined to not have occurred.

2.5.1 Actual

This is the number of offenses that occurred, simply a count of the number of crimes that month. For example if 10 people are murdered in a city the number of "actual murders" would be 10.

2.5.2 Total Cleared

A crime is cleared when an offender is arrested or when the case is considered cleared by exceptional means. When a single offender for a crime is arrested, that crime is considered cleared. If multiple people committed a crime, only a single person must be arrested for it to be cleared, and as the UCR data is at the offense level, making multiple arrests for an incident only counts as one incident cleared. So if 10 people committed a murder and all 10 were arrested, it would report one murder cleared not 10. If only one of these people are arrested it would still report one murder cleared - the UCR does not even say how many people commit a crime.

A crime is considered exceptionally cleared if the police can identify the offender, have enough evidence to arrest the offender, know where the offender is, but is unable to arrest them. Some examples of this are the death of the offender or when the victim refuses to cooperate in the case.

Unfortunately this data does not differentiate between clearances by arrest or by exceptional means. For a comprehensive report on how this variable can be exploited to exaggerate clearance rates, see this report by ProPublica on exceptional clearances with rape cases.

2.5.3 Cleared Where All Offenders Are Under 18

This variable is very similar to Total Cleared except is only for offenses in which **every** offender is younger than age 18.

2.5.4 Unfounded

An unfounded crime is one in which a police investigation has determined that the reported crime did not actually happen. For example if the police

are called to a possible burglary but later find out that a burglary did not occur, they would put it down as 1 unfounded burglary. This is based on police investigation rather than the decision of any other party such as a coroner, judge, jury, or prosecutor.

2.6 Number of months reported

UCR data is reported monthly though even agencies that decide to report their data may not do so every month. As we don't want to compare an agency which reports 12 months to one that reports fewer, the variable *number_of_months_reported* is way keep only agencies that report 12 months, or deal with those that report fewer.

From our `table()` output it seems that when agencies do report, they tend to do so for all 12 months of the year. However, this variable is seriously flawed, and its name is quite misleading. In reality this variable is actually just whichever the last month reported was. If an agency reported every month of the year, meaning December is the last month, they would have a value of 12. If the agency **only** reported in December, they would also have a value of 12. While there are ways in the monthly data to measure actual number of months reported, these ways are also flawed. So be cautious about this data and particularly the value of this variable.

Chapter 3

Arrests by Age, Sex, and Race

3.1 A brief history of the data

3.1.1 Changes in definitions

3.2 What does the data look like?

3.2.1 Raw data

3.2.2 Cleaned data

3.3 What variables are in the data?

3.3.1 Key variables

3.3.2 Known issues with the data

3.4 Final thoughts

Chapter 4

Law Enforcement Officers Killed and Assaulted (LEOKA)

4.1 A brief history of the data

4.1.1 Changes in definitions

4.2 What does the data look like?

4.2.1 Raw data

4.2.2 Cleaned data

4.3 What variables are in the data?

4.3.1 Key variables

4.3.2 Known issues with the data

4.4 Final thoughts

Chapter 5

Supplementary Homicide Reports (SHR)

5.1 A brief history of the data

5.1.1 Changes in definitions

5.2 What does the data look like?

5.2.1 Raw data

5.2.2 Cleaned data

5.3 What variables are in the data?

5.3.1 Key variables

5.3.2 Known issues with the data

5.4 Final thoughts

Chapter 6

Hate Crime Data

6.1 A brief history of the data

6.1.1 Changes in definitions

6.2 What does the data look like?

6.2.1 Raw data

6.2.2 Cleaned data

6.3 What variables are in the data?

6.3.1 Key variables

6.3.2 Known issues with the data

6.4 Final thoughts

Chapter 7

Property Stolen and Recovered (Supplement to Return A)

7.1 A brief history of the data

7.1.1 Changes in definitions

7.2 What does the data look like?

7.2.1 Raw data

7.2.2 Cleaned data

7.3 What variables are in the data?

7.3.1 Key variables

7.3.2 Known issues with the data

7.4 Final thoughts

Chapter 8

County-Level Detailed Arrest and Offense Data

8.1 A brief history of the data

8.1.1 Changes in definitions

8.2 What does the data look like?

8.2.1 Raw data

8.2.2 Cleaned data

8.3 What variables are in the data?

8.3.1 Key variables

8.3.2 Known issues with the data

8.4 Final thoughts